

WHAT IS CLAIMED IS:

1. A method for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method comprises:

providing for each of the plurality of genes a parameter that contains information concerning differences in the associated values of that gene among the sets;

adjusting the parameters of the plurality of genes so that the parameters are substantially independent of scatter values or average associated values of the genes over the sets;

deriving an observed value and an expected value of the adjusted parameter for each gene from the sets of associated values; and

comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance among the sets.

2. The method of claim 1, wherein said adjusting includes:

dividing the scatter values or average associated values of the genes into subsets each having a similar range of values, and calculating the standard deviation of each of the parameters within each subset;

altering the parameters until a coefficient of variation of the standard deviations of the parameters among the subsets is minimized.

3. The method of claim 1, further comprising obtaining said sets of associated values from multiple measurements of the plurality of genes, or values derived therefrom.

4. The method of claim 1, wherein said sets of associated values represent gene expression or number of gene copies or levels of protein encoded by the genes.

5. The method of claim 1, wherein said sets of associated values include calculated or predicted values.

6. The method of claim 1, wherein said providing includes calculating a difference value between an associated value of each gene in a first of the sets or a value

derived therefrom and an associated value of that gene in a second of the sets or a value derived therefrom; wherein the parameter is a function of the difference value of that gene.

7. The method of claim 6, wherein said providing further includes:

- 5 generating for each of the plurality of genes a scatter value that quantifies variation in the associated values of that gene within the first and second sets; and wherein said parameter is a function of the scatter value and of the difference value, said parameter defining a relative difference value of that gene.

8. The method of claim 7, wherein said generating employs the following equation:

$$s(i) = (\{1/a\} \{ \sum_m [x_m(i) - \underline{x}_I(i)]^2 + \sum_n [x_n(i) - \underline{x}_U(i)]^2 \})^{1/2}$$

where gene (i) has associated values $x_I(i)$ and $x_U(i)$ in Ith and Uth states respectively in the first and second sets of associated values, I and U being positive integers; \sum_m and \sum_n are sums over associated values of gene (i) in states I in the first set and in states U in the second set respectively, where $s(i)$ is the scatter value of gene (i), and a is a constant.

9. The method of claim 8, wherein said calculating calculates the parameter $d(i)$ from the following equation:

$$d(i) = [\underline{x}_I(i) - \underline{x}_U(i)]/[s(i) + s_0]$$

where s_0 is a constant, and $\underline{x}_I(i)$ and $\underline{x}_U(i)$ are the average values of $x_I(i)$ and $x_U(i)$ respectively in the first and second sets of associated values.

10. The method of claim 9, further comprising:

dividing the scatter values or average associated values of the genes into subsets each having a similar range of values, and calculating the standard deviation of each of the parameters within each subset; and

altering value of s_0 until a coefficient of variation of the standard deviations of the parameters among the subsets is minimized.

11. The method of claim 1, wherein said associated values of the genes are correlated with another variable so that each of said associated values has a corresponding

value of the variable, and wherein the parameter is provided using a Pearson correlation coefficient related to a weighted difference between each of the associated values and an average associated value, the variance of the associated values and the variance of the variable, said difference weighted by deviation of the corresponding value of the variable of such associated value from its average value.

12. The method of claim 11, wherein said variable is continuous.

13. The method of claim 12, wherein said variable is time.

14. The method of claim 11, wherein the parameter is selected using the Pearson correlation coefficient and a quantity s_0 that has a value adjusted as follows:

10 dividing the scatter values or average associated values of the genes into subsets each having a similar range of values, and calculating the standard deviation of each of the parameters within each subset; and

altering value of s_0 until a coefficient of variation of the standard deviations of the parameters among the subsets is minimized.

15 15. The method of claim 11, the number of sets of associated values being k , k being a positive integer, wherein said Pearson correlation coefficient $r(i)$ is given by:

$$r(i) = \sum_k [(x_k(i) - \bar{x}(i))[(y_k - \bar{y})]] / \sqrt{\sum_k (x_k(i) - \bar{x}(i))^2 \sum_k (y_k - \bar{y})^2}$$

where $x_k(i)$ is the associated value of gene (i) in the k th set of associated values, $\bar{x}(i)$ the average of the associated values of gene (i) in all the sets, y_k the value of the variable corresponding to $x_k(i)$, \bar{y} the average value of y_k in all the sets, and \sum_k is a sum over all values of k .

16. The method of claim 1, wherein the associated values in each set are classified into two or more subsets with values in each subset having a correlation with one another, and wherein the parameter is selected using a quantity related to variances between the associated values in the subsets of the sets and the variances of the associated values within each subset of the sets.

17. The method of claim 16, wherein the quantity relates to the sum of variances between the associated values in the subsets of the sets and the sum of variances of the associated values within each subset of the sets.

18. The method of claim 17, wherein the parameter is selected using the Fisher discriminant and a quantity s_0 having a value which has been adjusted as follows:

dividing the scatter values or average associated values of the genes into subsets each having a similar range of values, and calculating the standard deviation of each of the parameters within each subset; and

altering value of s_0 until a coefficient of variation of the standard deviations of the parameters among the subsets is minimized.

19. The method of claim 18, wherein the number of subsets of associated values of such set being k , k being a positive integer, and the Fisher discriminant $F(i)$ is given by:

$$F(i) = \sum_k n_k [x_k(i) - \bar{x}(i)]^2 / \sum_k \sum_j [x_j(i) - \bar{x}_k(i)]^2$$

where $x_k(i)$ is an associated value of gene (i) in the k th subset of associated values, $\bar{x}_k(i)$ the average of the associated values of gene (i) in the k th subset, $\bar{x}(i)$ the average value of the associated values of gene (i) in all of the subsets, n_k the number of associated values in the k th set, \sum_j a sum over all the associated values of gene (i) in the k th subset, and \sum_k a sum of the associated values of gene (i) over all of the subsets.

20. The method of claim 1, the sets of associated values referred to as original sets, wherein said deriving includes deriving said expected value by:

permuting, for each of the plurality of genes, the associated values for such gene in the original sets to arrive at a number of different permutations;

classifying the associated values in each permutation of each gene into corresponding permuted sets that are different from the original sets; and

supplying for each permutation a parameter value of each of the genes derived from an associated value of such gene in each of the corresponding permuted sets for such permutation or values derived therefrom.

21. The method of claim 20, wherein said associated values of the genes are correlated with another variable so that each of said associated values has an associated

value of the variable, wherein the permuting permutes the associated values so that at least each of some of the associated values has a different associated variable.

22. The method of claim 21, wherein the associated values are classified into two or more subsets with values in each subset having a correlation with one another,
5 wherein the permuting permutes the associated values so that at least each of some of the associated values is in a subset different from the subset it is classified into.

23. A method for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein the associated values correlate with
10 patient survival time, and wherein the associated values of the genes are obtained from a number of data sources, said method comprising:

- defining pairs of death and risk sets, each pair having a corresponding patient death time, where the death set of such pair includes associated values corresponding to the death time of such pair and the risk set of such pair includes associated values
15 corresponding to times occurring after the death time of such pair;

providing for each of the plurality of genes a parameter that contains information concerning differences in the associated values of that gene among the sets;

deriving an observed value and an expected value of the parameter for each gene from the sets of associated values; and

- 20 comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance.

24. The method of claim 23, wherein said providing provides said parameter as a function of weighted differences between the average associated values of the death and risk sets of the pairs, and of weighted variances within the risk sets.

25

25. The method of claim 24, wherein said providing provides for gene (*i*) said parameter by means of $r(i)$ and $s(i)$ given by the following:

$$r(i) = \sum_k d_k [x_k^*(i) - \underline{x}_k(i)]$$

$$s(i) = \{\sum_k (d_k/m_k) \sum_{j \in R(k)} [x_j(i) - \underline{x}_k(i)]^2\}^{1/2}$$

- 30 where there are K unique death times z_1, z_2, \dots, z_K ;

$D(k)$, for $k = 1, \dots, K$, are death sets defined by $D(k) = \{i : t_i = z_k\}$;

$R(k)$ are risk sets defined by $R(k) = \{i : t_i \geq z_k\}$;

m_k is number of patients in $R(k)$;

d_k is number of patient deaths at time z_k ;

an average expression of gene (i) in death set $D(k)$ is given by:

5 $x_k^*(i) = \sum_{j \in D(k)} x_j(i)/d_k$; and

an average expression of gene (i) in risk set $R(k)$ is given by:

$x_k(i) = \sum_{j \in R(k)} x_j(i)/m_k$.

26. The method of claim 24, wherein said providing provides said parameter
10 by means of $r(i)$ and $s(i)$ given by the following: $r(i)/[s(i)+s_0]$, where s_0 is a constant.

27. The method of claim 24, further comprising:

dividing the scatter values or average associated values of the genes into subsets
each having a similar range of values, and calculating the standard deviation of each of
15 the parameters within each subset; and

altering value of s_0 until a coefficient of variation of the standard deviations of the
parameters among the subsets is minimized.

28. A method for analyzing a plurality of original sets of values associated
with a plurality of genes to identify genes whose associated values differ by an amount of
20 statistical significance among the sets, wherein each of the sets of associated values of the
genes is obtained from one of a number of data sources, wherein the method comprises:

calculating for each gene a value for a statistical parameter indicating differences
between associated values of such gene among the original sets;

ranking the values of the parameter of the genes;

25 providing an expected value of such parameter for each rank, wherein said
providing includes permuting the associated values in the original sets to arrive at sets
different from the original sets for each permutation, deriving a value of such parameter
for each permutation, and ranking such values; and

30 comparing the calculated and expected values for the parameter of the same rank
to identify genes whose associated values differ by an amount of statistical significance
among the sets.

29. The method of claim 28, wherein said providing comprises:

for each permutation, deriving a value of the parameter for each gene and ranking the genes by their associated parameter values; and

determining the expected value of such parameter for each rank by computing an average value of the parameter of all the permutations having such rank.

5

30. The method of claim 29, wherein said comparing comprises identifying a gene as one whose associated values differ by an amount of statistical significance among the sets when the difference for such gene between the calculated value of the parameter of a rank and the expected value of such parameter of the same rank exceeds a threshold.

10

31. The method of claim 29, wherein said method further comprises identifying a lowest rank gene whose parameter value derived for a permutation is positive and exceeds a first threshold, setting such parameter value as a second threshold, comparing the derived parameter values of other genes for permutations to the second threshold and calling each gene whose derived parameter value exceeds the second threshold as a gene whose associated values are falsely identified to differ by an amount of statistical significance among the sets.

15

32. The method of claim 29, wherein said method further comprises identifying a lowest rank gene whose parameter value derived for a permutation is negative and less than a first threshold, setting such parameter value as a second threshold, comparing the derived parameter values of other genes for permutations to the second threshold and calling each gene whose derived parameter value is less than the second threshold as a gene whose associated values are falsely identified to differ by an amount of statistical significance among the sets.

20

33. The method of claim 28, wherein the sets of associated values in each permutation contains approximately an equal number of associated values from each of the original sets of associated values.

25

34. A method for analyzing a plurality of original sets of values associated with a plurality of genes to identify genes whose associated values are falsely identified to differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method comprises:

30

defining for each gene a statistical parameter indicating differences between associated values of such gene among the original sets;

providing an expected value of such parameter for each gene, wherein said providing includes permuting the associated values in the sets to arrive at sets different from the original sets for each permutation, deriving a value of such parameter for each permutation, and ranking such values;

deriving for each gene a value for the parameter for each permutation and ranking the genes by their derived parameter values;

finding a lowest rank gene whose derived parameter value extends beyond a first threshold; and

comparing the derived parameter values of other genes for permutations to the second threshold and calling each gene whose derived parameter value extends beyond the second threshold as a gene whose associated values are falsely identified to differ by an amount of statistical significance among the sets.

35. A method for reducing statistical error of a set of associated values of genes, wherein the method comprises:

providing a set of associated values of each gene; and

processing said set of associated values of that gene using a smooth weighting function to yield a representative value for that gene.

36. The method of claim 35, wherein said processing uses a Gaussian weighting function.

37. A method for comparing sets of associated values of genes, which comprises:

providing sets of associated values of each gene;

processing said sets of associated values of that gene using a smooth weighting function to obtain a representative value for that gene from each of the sets; and

comparing representative values for that gene for the sets.

38. The method of claim 37, wherein said providing includes calculating a difference PM-MM of a probe pair of a microarray.

39. A method for comparing a first and a second set of associated values of genes, which comprises:

providing odd root values of the values in the first set, and odd root values of the values in the second set; and

5 comparing the odd root values of the values in the first set and the odd root values of the values in the second sets.

40. The method of claim 39, wherein said providing provides the cube or fifth root values of the values in the first or second sets.

10 41. The method of claim 40, wherein said representing includes scaling the odd root values along the two axes, and wherein said method further comprises providing a best fit curve for the odd root values of the first and second set in the plot.

42. The method of claim 39, wherein said comparing includes representing the odd root values of the values in the first set along a first axis of a two-dimensional plot and the odd root values of the values in the second set along a second axis of the plot.

15 43. The method of claim 39, wherein said odd root values provided and compared includes values derived from positive and negative associated values.

44. A computer readable storage device embodying a program of instructions executable by a computer to perform a method for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method comprises:

20

providing for each of the plurality of genes a parameter that contains information concerning differences in the associated values of that gene among the sets;

25 adjusting the parameters of the plurality of genes so that the parameters are substantially independent of scatter values or average associated values of the genes over the sets;

deriving an observed value and an expected value of the adjusted parameter for each gene from the sets of associated values; and

30 comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance among the sets.

45. A computer readable storage device embodying a program of instructions executable by a computer to perform a method for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein the associated values correlate with patient survival time, and wherein the associated values of the genes are obtained from a number of data sources, said method comprising:

defining pairs of death and risk sets, each pair having a corresponding patient death time, where the death set of such pair includes associated values corresponding to the death time of such pair and the risk set of such pair includes associated values corresponding to times occurring after the death time of such pair;

providing for each of the plurality of genes a parameter that contains information concerning differences in the associated values of that gene among the sets;

deriving an observed value and an expected value of the parameter for each gene from the sets of associated values; and

comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance.

46. A computer readable storage device embodying a program of instructions executable by a computer to perform a method for analyzing a plurality of original sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method comprises:

calculating for each gene a value for a statistical parameter indicating differences between associated values of such gene among the original sets;

ranking the values of the parameter of the genes;

providing an expected value of such parameter for each rank, wherein said providing includes permuting the associated values in the original sets to arrive at sets different from the original sets for each permutation, deriving a value of such parameter for each permutation, and ranking such values; and

comparing the calculated and expected values for the parameter of the same rank to identify genes whose associated values differ by an amount of statistical significance among the sets.

47. A computer readable storage device embodying a program of instructions executable by a computer to perform a method for analyzing a plurality of original sets of values associated with a plurality of genes to identify genes whose associated values are
5 falsely identified to differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method comprises:

defining for each gene a statistical parameter indicating differences between associated values of such gene among the original sets;
10 providing an expected value of such parameter for each gene, wherein said providing includes permuting the associated values in the sets to arrive at sets different from the original sets for each permutation, deriving a value of such parameter for each permutation, and ranking such values;
deriving for each gene a value for the parameter for each permutation and ranking
15 the genes by their derived parameter values;
finding a lowest rank gene whose derived parameter value extends beyond a first threshold; and
comparing the derived parameter values of other genes for permutations to the second threshold and calling each gene whose derived parameter value extends beyond
20 the second threshold as a gene whose associated values are falsely identified to differ by an amount of statistical significance among the sets.

48. A computer readable storage device embodying a program of instructions executable by a computer to perform a method for reducing statistical error of a set of associated values of genes, wherein the method comprises:
25 providing a set of associated values of each gene; and
processing said set of associated values of that gene using a smooth weighting function to yield a representative value for that gene.

49. A computer readable storage device embodying a program of instructions executable by a computer to perform a method for comparing sets of associated values of
30 genes, which comprises:
providing sets of associated values of each gene;

processing said sets of associated values of that gene using a smooth weighting function to obtain a representative value for that gene from each of the sets; and comparing representative values for that gene for the sets.

50. A computer readable storage device embodying a program of instructions executable by a computer to perform a method for comparing a first and a second set of associated values of genes, which comprises:

providing odd root values of the values in the first set, and odd root values of the values in the second set; and

comparing the odd root values of the values in the first set and the odd root values of the values in the second sets.

51. A method for transmitting a program of instructions executable by a computer to perform a method for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method comprises:

causing a program of instructions to be transmitted to a client device, thereby enabling the client device to perform, by means of such program, the following process:

providing for each of the plurality of genes a parameter that contains information concerning differences in the associated values of that gene among the sets;

adjusting the parameters of the plurality of genes so that the parameters are substantially independent of scatter values or average associated values of the genes over the sets;

deriving an observed value and an expected value of the adjusted parameter for each gene from the sets of associated values; and

comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance among the sets.

52. A method for transmitting a program of instructions executable by a computer to perform a method for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein the associated values correlate with

patient survival time, and wherein the associated values of the genes are obtained from a number of data sources, said method comprising:

causing a program of instructions to be transmitted to a client device, thereby enabling the client device to perform, by means of such program, the following process:

5 defining pairs of death and risk sets, each pair having a corresponding patient death time, where the death set of such pair includes associated values corresponding to the death time of such pair and the risk set of such pair includes associated values corresponding to times occurring after the death time of such pair;

providing for each of the plurality of genes a parameter that contains information
10 concerning differences in the associated values of that gene among the sets;

deriving an observed value and an expected value of the parameter for each gene from the sets of associated values; and

comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance.

15 53. A method for transmitting a program of instructions executable by a computer to perform a method for analyzing a plurality of original sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method
20 comprises:

causing a program of instructions to be transmitted to a client device, thereby enabling the client device to perform, by means of such program, the following process:

calculating for each gene a value for a statistical parameter indicating differences between associated values of such gene among the original sets;

25 ranking the values of the parameter of the genes;

providing an expected value of such parameter for each rank, wherein said providing includes permuting the associated values in the original sets to arrive at sets different from the original sets for each permutation, deriving a value of such parameter for each permutation, and ranking such values; and

30 comparing the calculated and expected values for the parameter of the same rank to identify genes whose associated values differ by an amount of statistical significance among the sets.

54. A method for transmitting a program of instructions executable by a computer to perform a method for analyzing a plurality of original sets of values associated with a plurality of genes to identify genes whose associated values are falsely identified to differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the method comprises:

causing a program of instructions to be transmitted to a client device, thereby enabling the client device to perform, by means of such program, the following process:

defining for each gene a statistical parameter indicating differences between associated values of such gene among the original sets;

providing an expected value of such parameter for each gene, wherein said providing includes permuting the associated values in the sets to arrive at sets different from the original sets for each permutation, deriving a value of such parameter for each permutation, and ranking such values;

deriving for each gene a value for the parameter for each permutation and ranking the genes by their derived parameter values;

finding a lowest rank gene whose derived parameter value extends beyond a first threshold; and

comparing the derived parameter values of other genes for permutations to the second threshold and calling each gene whose derived parameter value extends beyond the second threshold as a gene whose associated values are falsely identified to differ by an amount of statistical significance among the sets.

55. A method for transmitting a program of instructions executable by a computer to perform a method for reducing statistical error of a set of associated values of genes, wherein the method comprises:

causing a program of instructions to be transmitted to a client device, thereby enabling the client device to perform, by means of such program, the following process:

providing a set of associated values of each gene; and

processing said set of associated values of that gene using a smooth weighting function to yield a representative value for that gene.

56. A method for transmitting a program of instructions executable by a computer to perform a method for comparing sets of associated values of genes, which comprises:

- causing a program of instructions to be transmitted to a client device, thereby
- 5 enabling the client device to perform, by means of such program, the following process:
 - providing sets of associated values of each gene;
 - processing said sets of associated values of that gene using a smooth weighting function to obtain a representative value for that gene from each of the sets; and
 - comparing representative values for that gene for the sets.

10 57. A method for transmitting a program of instructions executable by a computer to perform a method for comparing a first and a second set of associated values of genes, which comprises:

- causing a program of instructions to be transmitted to a client device, thereby
- enabling the client device to perform, by means of such program, the following process:
 - 15 providing odd root values of the values in the first set, and odd root values of the values in the second set; and
 - comparing the odd root values of the values in the first set and the odd root values of the values in the second sets.

20 58. A computer system for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the system comprises:

- one or more computers;
- one or more computer programs running on the computer(s), performing the
- 25 following:
 - providing for each of the plurality of genes a parameter that contains information concerning differences in the associated values of that gene among the sets;
 - adjusting the parameters of the plurality of genes so that the parameters are substantially independent of scatter values or average associated values of the genes over
 - 30 the sets;
 - deriving an observed value and an expected value of the adjusted parameter for each gene from the sets of associated values; and

comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance among the sets.

59. A computer system for analyzing a plurality of sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein the associated values correlate with patient survival time, and wherein the associated values of the genes are obtained from a number of data sources, said system comprising:

one or more computers;
10 one or more computer programs running on the computer(s), performing the following:

defining pairs of death and risk sets, each pair having a corresponding patient death time, where the death set of such pair includes associated values corresponding to the death time of such pair and the risk set of such pair includes associated values
15 corresponding to times occurring after the death time of such pair;

providing for each of the plurality of genes a parameter that contains information concerning differences in the associated values of that gene among the sets;

deriving an observed value and an expected value of the parameter for each gene from the sets of associated values; and

20 comparing the observed and expected values of the parameter to identify genes whose associated values differ by an amount of statistical significance.

60. A computer system for analyzing a plurality of original sets of values associated with a plurality of genes to identify genes whose associated values differ by an amount of statistical significance among the sets, wherein each of the sets of associated
25 values of the genes is obtained from one of a number of data sources, wherein the system comprises:

one or more computers;

one or more computer programs running on the computer(s), performing the following:

30 calculating for each gene a value for a statistical parameter indicating differences between associated values of such gene among the original sets;

ranking the values of the parameter of the genes;

providing an expected value of such parameter for each rank, wherein said providing includes permuting the associated values in the original sets to arrive at sets different from the original sets for each permutation, deriving a value of such parameter for each permutation, and ranking such values; and

- 5 comparing the calculated and expected values for the parameter of the same rank to identify genes whose associated values differ by an amount of statistical significance among the sets.

61. A computer system for analyzing a plurality of original sets of values
10 associated with a plurality of genes to identify genes whose associated values are falsely identified to differ by an amount of statistical significance among the sets, wherein each of the sets of associated values of the genes is obtained from one of a number of data sources, wherein the system comprises:

one or more computers;

- 15 one or more computer programs running on the computer(s), performing the following:

defining for each gene a statistical parameter indicating differences between associated values of such gene among the original sets;

- providing an expected value of such parameter for each gene, wherein said
20 providing includes permuting the associated values in the sets to arrive at sets different from the original sets for each permutation, deriving a value of such parameter for each permutation, and ranking such values;

deriving for each gene a value for the parameter for each permutation and ranking the genes by their derived parameter values;

- 25 finding a lowest rank gene whose derived parameter value extends beyond a first threshold; and

- comparing the derived parameter values of other genes for permutations to the second threshold and calling each gene whose derived parameter value extends beyond the second threshold as a gene whose associated values are falsely identified to differ by
30 an amount of statistical significance among the sets.

62. A computer system for reducing statistical error of a set of associated values of genes, wherein the system comprises:

one or more computers;

one or more computer programs running on the computer(s), performing the following:

providing a set of associated values of each gene; and

5 processing said set of associated values of that gene using a smooth weighting function to yield a representative value for that gene.

63. A computer system for comparing sets of associated values of genes, which comprises:

one or more computers;

10 one or more computer programs running on the computer(s), performing the following:

providing sets of associated values of each gene;

processing said sets of associated values of that gene using a smooth weighting function to obtain a representative value for that gene from each of the sets; and

comparing representative values for that gene for the sets.

15

64. A computer system for comparing a first and a second set of associated values of genes comprising

one or more computers;

20 one or more computer programs running on the computer(s), performing the following:

providing odd root values of the values in the first set, and odd root values of the values in the second set; and

comparing the odd root values of the values in the first set and the odd root values of the values in the second sets.

25